1 Overview

This test is inspired by a recent EPIC research project. Your task is to perform basic data cleaning, prepare a final dataset for analysis, provide short answers to prompts, and create publication-quality figures and tables.

This task should take **five hours or less, but you will have 48 hours total to complete it**. The goal of the task is to give you an opportunity to demonstrate your coding competency and your conceptual understanding of empirical economics research. A perfect grade is not a prerequisite for consideration for the position.

For the coding portion of this test, we will accept code in Python, R, or STATA. You may consult any pre-existing online programming resources, but you may not ask other people for help. If you find any of these instructions to be confusing, please proceed in a way that you find relevant and reasonable, and list your assumptions in your writeup. Once you have completed the sections below, please submit the following in a .zip file:

- Well-commented code in the language of your choice (.do files for Stata, .R or .Rmd files for R, .py or .ipynb files for Python, etc.),
- The final dataset from Section 2,
- The final graphs and tables from Section 3,
- A short document answering the questions from Sections 2-4

2 Data Cleaning

The central task of this section is to merge `production data` and `price data` into one dataset. Both data are sourced from the US Department of Agriculture. Both datasets contain annual data from 1990 to 2018 A brief introduction on the datasets:

- **Barley_production.csv** lists the barley production in bushels by agricultural district. The agricultural district is an administrative division between the county and state levels.
- **Barley_price.csv** lists the mean price received by farmers per bushel of barley by state. Ignore the distinction between the marketing year and the calendar year.

We want the final dataset to be:

- a panel with three dimensions: year, agricultural district, state
- in each row it contains: barley production and price

3 Data Exploration

The data exploration in this section would be conducted at **state-year** level.

3.1 Time Series Plot: Price

For each year, compute the weighted average of price over all states, where each state's weight is its production in bushels in that year. Then, plot this weighted average over the time period from 1990 to 2018.

3.2 Time Series Plot: Production

Find the top 3 states in terms of barley production in 2018. Plot the time series of production for these 3 states in the same plot, over the period from 1990 to 2018. Scale the production variable so that it is in millions of bushels.

3.3 Summary Table

Create a summary table where the rows are specific states (Idaho, Minnesota, Montana, North Dakota, and Wyoming) and the columns are decades (1990-1999, 2000-2009, and 2010-2018). The elements of the table are mean annual state-level production, by decade and state. Scale the production variable so that it is in millions of bushels.

4 Short Answer

Our goal is to estimate the sensitivity of US farmers' barley production to barley price, using the provided data, at the level of agricultural district by year.

- First write down a regression equation of a linear model of production on a constant and price. We want the coefficient on price to have the interpretation of an elasticity. Ensure that the terms are properly indexed. Report the results of this regression, and interpret the coefficient on price.
- What variables do you think we should control for? Choose two and explain why they might help us identify the coefficient on price. These variables need not to be in the original dataset.
- Price is an endogenous variable in our model. Provide examples of two different types of endogeneity that could bias our estimated coefficient on price.

- We can somewhat mitigate this problem by including year and state fixed effects. Run this regression on the provided data and report the estimated coefficient on price, along with its standard error. Justify the method you used to adjust the standard error. Is this coefficient causally interpretable?

- Which potential sources of price endogeneity does adding fixed effects address? Discuss the difference (if any) in results with the results above. Which sources might still remain? Make sure to provide concrete examples.

- To address any remaining sources of price endogeneity, suppose that we use the transportation cost between location of barley production and its market as an instrument for price. Explain why you think this is or is not a valid instrument. What challenges could arise when using this instrument in practice?

- Suppose one of the other research assistants accidentally deletes 10% of the observations of the barley production variable. How would you expect dropping these observations to change the estimated coefficient on price and its standard error if the deletions are at random? What if the deletions are not at random?